

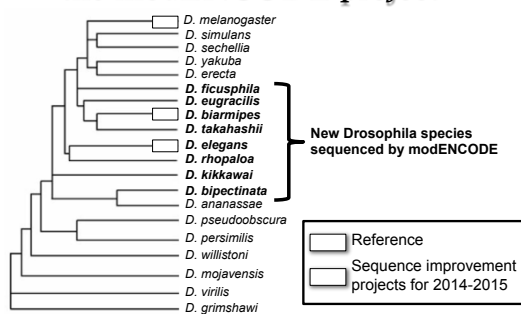
Overview of the *Drosophila* modENCODE hybrid assemblies

Wilson Leung 01/2014

Agenda

- ⊗ Overview of the modENCODE species
- ⊗ Problems with the version 1 genome assembly
- ⊗ Improvements in the version 2 genome assembly
- ⊗ Pipeline used to create the sequence improvement projects
- ⊗ Sequence improvement goals for hybrid assemblies

New *Drosophila* species sequenced by the modENCODE project



The *Drosophila* modENCODE sequencing project

- ⊗ Selected 8 additional *Drosophila* species for sequencing based on the ideal evolutionary distances for motif finding
- ⊗ Sequenced and assembled by the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) in 2011
- ⊗ BCM-HGSC also produced RNA-seq data for each newly sequenced species
 - ⊗ Adult males, adult females, and mixed embryos
- ⊗ Initial assembly based on only 454 reads
 - ⊗ ~45x coverage
 - ⊗ 15x unpaired reads with 454 XLR
 - ⊗ 30x paired end sequencing of 3kb and 8kb inserts

Version 1 assembly statistics

Species	Total Length	Gap Length	Scaffold Count	Scaffold N50
<i>D. ficusphila</i>	152,437,402	1,391,609	5,761	1,050,437
<i>D. eugracilis</i>	156,925,476	612,291	4,947	976,726
<i>D. biarmipes</i>	169,375,387	787,648	5,528	3,385,622
<i>D. takahashii</i>	182,069,776	1,069,231	5,734	387,609
<i>D. elegans</i>	171,239,642	722,655	5,429	1,713,827
<i>D. rhopaloa</i>	197,384,455	3,482,582	22,825	45,465
<i>D. kikkawai</i>	164,317,633	835,259	5,142	903,526
<i>D. bipectinata</i>	167,246,727	852,176	5,505	657,776

- ⊗ *D. biarmipes* has the highest N50 among the 8 species
 - ⊗ N50 = Total size of scaffolds this size or larger that account for half of the total length of the entire assembly
- ⊗ Quality of the *D. rhopaloa* assembly is much lower than the other 7 genomes

Error profile for 454 reads

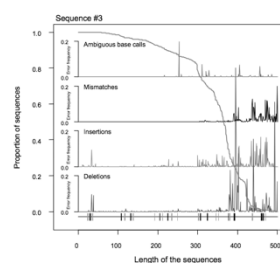


Figure 1 Distribution of errors along sequences. The blue line indicates the proportion of generated sequences (y-axis) as a function of sequence position (x-axis), based on data obtained from the analysis of reference sequence #3. The error rate for each type
 Gilles *et al.* Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 2011, 12:245

- ⊗ Length of each 454 read depends on the number of cycles and base composition of the read
- ⊗ Generally trim the end of 454 reads because they are low quality
- ⊗ Most of the remaining errors are base insertions or deletions (indels)

Sequence improvement goals for hybrid assemblies

- ⊛ **Primary goal:** Correct errors within mononucleotide runs
 - ⊛ **Secondary goal:** Close gaps and correct regions with low consensus quality
 - ⊛ **Optional goal:** Identify regions with putative polymorphisms
- ⊛ Recommended protocols and training materials for improving hybrid assemblies available on the GEP web site under the [GEP Sequence Improvement Projects Issues](#) section

Questions?



<http://www.flickr.com/photos/omcoc/6751047205/sizes/l/>